From historic books to annotated XML: Building a large multilingual diachronic corpus

From Historic Books to Annotated XML: Building a Large Multilingual Diachronic Corpus

Magdalena Jitca, Rico Sennrich, Martin Volk
Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zürich
E-mail: mjitca, sennrich, volk @ifi.uzh.ch

Abstract

This paper introduces our approach towards annotating a large heritage corpus, which spans over 100 years of alpine literature. The corpus consists of over 16.000 articles from the yearbooks of the Swiss Alpine Club, 60% of which represent German texts, 38% French, 1% Italian and the remaining 1% Swiss German and Romansh. The present work describes the inherent difficulties in processing a multilingual corpus by referring to the most challenging annotation phases such as article identification, correction of optical character recognition (OCR) errors, tokenization, and language identification. The paper aims to raise awareness for the efforts in building and annotating multilingual corpora rather than to evaluate each individual annotation phase.

Keywords: multilingual corpora, cultural heritage, corpus annotation, text digitization

1. Introduction

In the project Text+Berg[1] we are digitizing publications of the Alpine clubs from various European countries, which consist mainly of reports on the following topics: mountain expeditions, the Alpine culture, the flora, fauna and geology of the mountains.

The resulting corpus is a valuable knowledge base to study the changes in all these areas. Moreover, it enables the quantitative analysis of diachronic language changes as well as the study of typical language structures, linguistic topoi, and figures of speech in the mountaineering domain.

This paper describes the particularities of our corpus and gives an overview of the annotation process. It presents the most interesting challenges that our multilingual corpus brought up, such as text structure identification, optical character recognition (OCR), tokenization, and language identification. We focus on how the multilingual nature of the text collection poses new problems in apparently trivial processing steps (e.g. tokenization).

[1] See www.textberg.ch

2. The Text+Berg Corpus

The focus of the Text+Berg project is to digitize the yearbooks of the Swiss Alpine Club from 1864 until today. The resulting corpus contains texts which focus on conquering and understanding the mountains and covers a wide variety of text genres such as expedition reports, (popular) scientific papers, book reviews, etc.

The corpus is multilingual and contains articles in German (some also in Swiss German), French, Italian and even Romansh. Initially, the yearbooks contained mostly German articles and few in French. Since 1957 the books appeared in parallel German and French versions (with some Italian articles), summing up to a total of 53 parallel editions German-French and 90 additional multilingual yearbooks. The corpus contains 16.000 articles, 60% of which represent German texts, 38% French, 1% Italian and the remaining 1% Swiss German and Romansh. This brings our corpus to 35,75 million words extracted from almost 87.000 book pages, 10% of which representing parallel texts. This feature of the corpus allows for interesting cross-language comparisons and has been used as training material for Statistical Machine Translation systems (Sennrich & Volk, 2010).

3. The Annotation Phases

This section introduces our pipeline for processing and annotating the Text+Berg corpus. More specifically, the input consists of HTML files containing the scanned yearbooks (for yearbooks in paper format), as they are exported by the OCR software. We work with two state-of-the-art OCR programs (Abbyy FineReader 7 and OmniPage 17) in order to convert the scan images into text and then export the files in HTML format. Our processing pipeline takes them through ten consecutive stages: 1) HTML cleanup, 2) structure reducing, 3) OCR merging, 4) article identification, 5) parallel book combination, 6) tokenization, 7) correction of OCR errors, 8) named entity recognition, 9) Part of Speech (POS) tagging and 10) additional lemmatization for German. The final output consists of XML documents which mark the article structure (title, author), as well as sentence boundaries, tokens, named entities (restricted to mountain, glacier and cabin names), POS tags and lemmas. Our document processing approach is similar to other annotation pipelines, such as GATE (Cunningham et al., 2002), but it is customized for our alpine corpus. In terms of space complexity, the annotated output files require almost three times more storage space than the input HTML files and 2,3 times more space than the tokenized XML files, respectively.

In the following subsections we expand on the processing stages that are especially challenging for a multilingual corpus.

3.1. Article Identification

The identification of articles in the text is performed during the fourth processing stage. The text is annotated conforming to an XML schema which marks the article boundaries (start, end), its title and author, paragraphs, page breaks, footnotes and captions. Some of the text structure information can be checked against the table of contents (ToC) and table of figures (where available), which are manually corrected in order to have a clean database of all articles in the corpus. Another relevant resource for the article boundary identification is the page mapping file that is automatically generated in the second stage, which relates the number printed on the original book page with the page number assigned during scanning. The process of matching entries from the table of contents to the article headers in the books is not trivial, as it requires that the article title, the author name(s) and the page number in the book are correctly recognized. We allow small variations and OCR errors, as long as they are below a specific threshold (usually a maximum deviation of 20% of characters is allowed). For example, the string *K/aIbard -Eine Reise in die Eiszeit.* will be considered a match for the ToC entry *Svalbard - Eine Reise in die Eiszeit*, although not all their characters coincide.

Proper text structuring relies on the accurate identification of layout elements such as article boundaries, graphics and captions, headers and footnotes. Over the 145 years the layout of the yearbooks has changed significantly. Therefore we had to adapt different processing steps for all the various designs. The particularities of these layouts have been discussed in (Volk et al., 2010a).

The yearbooks since 1996 are a collection of monthly editions and their pagination is no longer continuous (it starts over every month). This change affects the page mapping process, which performs well only when page numbers are monotonically increasing. Moreover, article boundaries are hard to determine when a single page contains several small articles and not all of them specify their author's name. These particularities are also reflected in the layout, as the header lines (where existing) no longer contain information about author or title, but about the article genre. Under these circumstances, we still achieved a percentage of 80% identified articles for these new yearbooks, a value comparable to the overall percentage of the corpus.

3.2. Correction of OCR Errors

The correction process aims to detect and overcome the errors introduced by the OCR systems and is carried out in two different stages of the annotation process. The first revision is done in the third stage (OCR merging), where the input is still raw text, with no additional information about either the structure or the language of the articles. At this stage we combine the output of our two OCR systems. The algorithm computes the alignments in a page-level comparison of the input files provided by each system and searches the Longest Common Subsequence in a n-character window. In case

of mismatch, the system disambiguates among the different candidates and selects the word with the highest probability in that context (computed based on the word's frequency in the Text+Berg corpus). The implemented algorithm and the evaluation results are thoroughly discussed in (Volk et al., 2010b).

OCR-merging is a worthwhile approach since there are many situations where one system can fix the other's errors. Our experience has shown that Abbyy FineReader performs the better OCR, with over 99% accuracy (Volk et al., 2010b). But there are also cases where it fails to provide the correct output, whereas OmniPage provides the right one. For example, the sequence *Cependant, les cartes disponibles sont* ***squvent*** *approximatives* (English: However, the available maps are often approximate) is provided by FineReader. The system has introduced the spelling mistake *squvent*, which doesn't appear in the output of the second system (here *souvent*). This triggers the replacement of the non-word *squvent* with the correct version *souvent*.

During the seventh annotation stage, after tokenization, we correct errors caused by graphemic similarities. The automatic correction is performed at the word-level by pattern matching over sequences of characters. In order to achieve this, we have compiled lists of common error patterns and their possible replacements. For example, a word-initial 'R' is often misinterpreted as 'K', resulting in words such as *Kedaktion* instead of *Redaktion* (English: editorial office). For each tentative replacement we check against the word frequency list in order to decide whether a candidate word appears in the corpus more frequently than the original or the other possible replacement candidates. In this case, *Redaktion* has 1127 occurrences in the corpus, whereas *Kedaktion* only 9. Reynaert (2008) describes a similar statistical approach for both historical and contemporary texts.

As the yearbooks until 1957 contained articles written in several languages, we have used a single word frequency dictionary for all of them (German, French and Italian). The dictionary has been built from the Text+Berg corpus and thus contains all the encountered word types and their corresponding frequencies, computed over the same corpus. The interesting aspect about this dictionary is its reliability, in spite of being trained with noisy data (text containing OCR-errors). Correctly spelled words will typically have a higher frequency than the ones containing OCR errors. The list contains predominantly German words due to the high percentage of German articles in the first 90 yearbooks, thus the frequency of German words is usually higher than that of French words. This can lead to wrong substitution choices, such as a German word in a French sentence (e.g. *Neu* (approx. 4400 hits) instead of *lieu* (approx. 3000 hits)). Therefore we have decided to create a separate frequency dictionary for French words, which is used only for the monolingual French editions.

3.3. Tokenization

In this stage the paragraphs of the text are split into sentences and words, respectively. Tokenization is considered to be a straightforward problem that can be solved by applying a simple strategy such as split on all non-alphanumeric characters (e.g. spaces, punctuation marks). Studies have shown, however, that this is not a trivial issue when dealing with hyphenated compound words or other combinations of letters and special characters (e.g. apostrophes, slashes, periods etc.). He and Kayaalp (2006) present a comparative study of several tokenizers for English, showing that their output varies widely even for the same input language. We would expect a similar performance from a general purpose tokenizer dealing with several languages.

We will exemplify the language-specific issues with the use of apostrophes. In many languages, they are used for contractions between different parts of speech, such as verb + personal pronoun *es* in German (e.g. *hab's* → *habe* + *es*) or determiner and noun in French or Italian (e.g. *l'abri* → *le* + *abri*). On the other hand, in old German written until 1900, like in modern English, it can also express possession (e.g. *Goldschmied's*, *Theobald's*, *Mozart's*). Under these circumstances, which is the desired tokenization, before or after the apostrophe? The answer is language-dependent and this underlies our approach towards tokenization.

We use a two-step tokenization and perform the language recognition in between. The advantage of this approach is that we can deliver a language-specific tokenization of any input text (given that it is written in the supported languages). In the first step we carry out a rough tokenization of the text and then identify sentence

boundaries. Once this is achieved, we can proceed to the language identification, which will be discussed in section 3.4.

Afterwards we do another round of tokenization focused on word-level, where the language-specific rules come into play. We have implemented a set of heuristic rules in order to deal with special characters in a multilingual context, such as abbreviations, apostrophes or hyphens. For example, each acronym whose letters are separated by periods (e.g. C.A.S. or A.A.C.Z.) is considered a single token, if it is listed in our abbreviations dictionary. A German apostrophe is split from the preceding word (e.g. *geht's* → *geht* + *'s*), whereas in French and Italian it remains with the first word (e.g. *dell'aqua* → *dell'* + *aqua*, *l'eau* → *l'* + *eau*). Besides, we have compiled a small set of French apostrophe words which shouldn't be separated at all (e.g. *aujourd'hui*).

Disambiguation for hyphens occurring in the middle of a word is performed by means of the general word frequency dictionary. For example, if *nordouest* has 14 hits and *nord-ouest* 957 hits, we conclude that the hyphen is part of the compound and thus *nord-ouest* should be regarded as a single token. On the other hand, hyphens marking line breaks may also appear in the middle, like in the word *rou-te*. In this case, the hyphenated word appears 3 times in the dictionary, whereas the one without, *route*, 6335 times. Therefore the hyphen will be removed from the word.

3.4. Language Identification

The accuracy of the language identification is crucial for the automatic text analysis performed during the annotation process, such as tokenization, part-of-speech tagging, lemmatization or named entity identification. Therefore we perform a fine-grained analysis, at sentence level. We work with a statistical language identifier[2] based on the approach presented in (Dunning, 1994). The module uses two classifiers: one to distinguish between German, French, English and Italian and another one in order to discriminate between Italian and Romansh. In case the identified language is German, a further analysis based on the frequency dictionary is being carried out in order to decide whether or not it is Swiss German (CH-DE). This dictionary contains frequently used Swiss German dialect words which do not have homographs in standard German. Whenever a sentence contains more than 10% dialect words from this list, the language of the sentence is set to CH-DE.

However, the statistical language identification is not reliable for very short sentences. In order to achieve higher accuracy, we apply the heuristic rule that only sentences longer than 40 characters are fed to the language identifier. All the others are assigned the language of the article, as it appears in the ToC. The correctness of this decision relies on the fact that all ToC files are proofed manually, so that we do not introduce noisy data.

Table 1 gives an overview of the distribution of the identified languages in the articles from the Text+Berg corpus. We present here only the composition of German and French articles, as they represent the great majority of our corpus (approximatively 98%). The values are not 100% accurate, as they are automatically computed by means of statistical methods. However, they mirror the global tendencies of the corpus that over 95% of the sentences in an article are in the language of the article, a conclusion which corresponds to our expectations. An interesting finding is the percentage variation of foreign sentences. For example, German sentences are two times more frequent in French articles than the French sentences in German articles (in percentage terms). One reason for this is the fact that some French articles are translated from German and preserve the original bibliographical references, captions or footnotes. Other sources of language mixture are quotations and direct speech, aspects which can be encountered in both German and French articles.

3.5. Linguistic Processing

In the last two annotation stages we perform some linguistic processing, namely lemmatization and part of speech tagging. The markup is done by the TreeTagger[3]. For our corpus, we have applied the standard configuration files for German, English and Italian. In the case of French we adopted a different approach, and we have trained our own parameter files based on the Le Monde-Treebank (Abeillé, 2003).

[2] http://search.cpan.org/dist/Lingua-Ident/Ident.pm

[3] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Article language	Number of sentences per language						
	de	en	fr	it	rm	ch-de	total
DE	1.166.141	1035	11.607	1481	1490	799	1.182.553
FR	12.392	607	670.599	1187	1277	2	686.064

Table 1: The language distribution of the sentences in the Text+Berg corpus

```
<book id="1901_mul">
  <article n="5">
    <tocEntry title="Altes und Neues aus dem Säntisgebiet"
              author="C. Egloff" lang="de" category="Freie Fahrten"/>
    <div>
      <s n="5-88" lang="de">
        <w n="5-88-1" pos="APPRART" lemma="im">Im</w>
        <w n="5-88-2" pos="NN" lemma="Eiltempi">Eiltempo</w>
        <w n="5-88-3" pos="VVFIN" lemma="gehen">ging</w>
        <w n="5-88-4" pos="PPER" lemma="es">'s</w>
        <w n="5-88-5" pos="PTKVZ" lemma="weiter">weiter</w>
        <w n="5-88-6" pos="$," lemma=",">,</w>
        <w n="5-88-7" pos="ADV" lemma="erst">erst</w>
        <w n="5-88-8" pos="ADJD" lemma="kletternd">kletternd</w>
        <w n="5-88-9" pos="$," lemma=",">,</w>
        <w n="5-88-10" pos="ADV" lemma="dann">dann</w>
        <w n="5-88-11" pos="APPR" lemma="über">über</w>
        <w n="5-88-12" pos="NN" lemma="Felstrümmer">Felstrümmer</w>
        <w n="5-88-13" pos="KON" lemma="und">und</w>
        <w n="5-88-14" pos="NN" lemma="Schneefeld">Schneefelder</w>
        <w n="5-88-15" pos="$." lemma=";">;</w>
      </s>
    </div>
```

Figure 1: An annotation snippet

Romansh is not yet supported due to the lack of a sufficiently large annotated corpus for training the corresponding parameter file. Figure 1 shows a sample output: an annotated sentence in XML format.

The TreeTagger assigns only lemmas for word forms that it knows (that have been encountered during the training). This results in a substantial number of word forms with unknown lemmas. Therefore we use an additional lemmatization tool, in order to increase the coverage of lemmatization. This approach has been implemented for German only because of its large number of compounds.

We use the system Gertwol[4] to insert missing German lemmas. Towards this goal we collect all word form types from the corpus and have Gertwol analyse them. If the TreeTagger does not assign a lemma to a word, whereas Gertwol provides an appropriate alternative, we choose the output of the latter system. This has resulted in approximately 700.000 additional lemmas, 80% percent of which represent noun lemmas, 15% adjectives and the remaining 5% other parts of speech.

After performing this step, the remaining unknown lemmas are mostly names and words containing OCR errors. We are interested in extending this strategy for French and Italian, in order to further increase the coverage of the annotation.

4. Tools for Accessing the Corpus

The Text+Berg corpus can be accessed through several search systems. For example, we have stored our annotated corpus in the Corpus Query Workbench (Christ, 1994), which allows us to browse it via a web interface[5]. The queries follow the POSIX EGREP syntax for regular expressions. The advantage of this system is that it provides more precise results than usual search engines (which perform a full text search) due to our detailed annotations. For example, it is possible to query for all mountain names ending in *horn* that were mentioned before 1900. Moreover, it is also possible to restrict queries to particular languages or POS tags.

In addition, we have built a tool for word alignment searches in our parallel corpus[6]. Given a German search term, the tool displays all hits in the German part of the corpus together with the corresponding French sentences with the aligned word(s) highlighted. Other than being a word alignment visualization tool, it also serves as bilingual concordance tool to find mountaineering terminology in usage examples. In this way it is easy to determine the appropriate translation for words like *Haken* (English: hook) or *Steigeisen* (English: crampon). Moreover, it enables a consistent view of the possible translations of ambiguous words as *Kiefer* (English: jaw, pine) or *Mönch* (English: monk, mountain name). Figure 2 depicts the output of the system for the word *Leiter*, which can either refer to leader or ladder.

[4]http://www2.lingsoft.fi/cgi-bin/gertwol

[5]Access to the CQW is password-protected. See http://www.textberg.ch/index.php?id=4&lang=en for registration.

[6]http://kitt.ifi.uzh.ch/kitt/alignsearch/

1959, article 8 Colin Wyatt: *Bergfahrt durch Nepal*	Eine zweite **Leiter** führte uns durch eine Deckenöffnung aufs Dach , von wo man das Dorf und das Tal überblickte .	Une autre **échelle** nous conduisit par un trou du plafond sur le toit en terrasse dominant le village et la vallée .
1959, article 41 G.O. Dyhrenfurth: *Himalaya-Chronik 1958*	**Leiter** war Gurdial Singh , Honorary Local Secretary des Himalayan Club in Dehra Dun .	**Le chef** en était Gourdial Singh , secrétaire local de l' Himalayan Club à Dehra Dun .
1984, article 4 Lorenz Seiler, Roland Radlinger: *Situationswahrnehmung und Angstentstehung im Bergsport*	Innerhalb der Gruppe nimmt **der Leiter** eine besondere Stellung ein :	A l' intérieur du groupe , **le moniteur** occupe une place particulière :
1985, article 14 Trevor Braham: *Himalaya-Chronik 1984*	Zwei der Briten , der **Leiter** und M. Fowler stiegen weiter bis ca. 7000 m auf , bevor sie umkehrten .	Deux des grimpeurs , le **chef de l' expédition** et M. Fowler , ont continué l' ascension jusqu' à 7000 mètres environ , où ils ont fait demi-tour .

Figure 2: Different translations of the German word *Leiter* in the Text+Berg corpus

5. Conclusion

In this paper we have given an overview of the annotation workflow of the Text+Berg corpus. The pipeline is capable of processing multilingual documents and dealing with both diachronic varieties in language and noisy data (OCR errors). The flexible architecture of the pipeline allows us to extend the corpus with more alpine literature and to process it in a similar manner, with little overhead.

We have provided insights into the multilingual challenges in the annotation process, such as OCR correction, tokenization or language identification. We intend to further reduce the number of OCR errors by launching a crowd correction wiki page, where the members of the Swiss Alpine Club will be able to correct such mistakes. Regarding linguistic processing, we will continue investing efforts in improving the quality of the existing annotation tools with language-specific resources (e.g. frequency dictionaries, additional lemmatizers). We will also work on improving the language models for Romansh and Swiss German dialects, in order to increase the reliability of the language identifier.

6. References

Abeillé, A., Clément, L., Toussenel, F. (2003): Building a Treebank for French. In Building and Using Parsed Corpora, Text, Speech and Language Technology(20), pp. 65–187.

Christ, O. (1994): The IMS Corpus Workbench Technical Manual. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Cunningham, H., Maynard, D., Bontcheva, K. (2002): GATE: A framework and graphical development environment for robust NLP tools and applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.

Dunning, T. (1994): Statistical identification of language. Technical Report MCCS-94-273, New Mexico State University.

He, Y., Kayaalp, M. (2006): A comparison of 13 tokenizers on MEDLINE. Technical Report LHNCBC-TR-2006-003, The Lister Hill National Center for Biomedical Communications.

Reynaert, M. (2008): Non-interactive OCR post-correction for giga-scale digitization projects. In A. Gelbukh (Ed.), Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, Lecture Notes in Computer Science. Berlin, Springer, pp. 617–630.

Sennrich, R., Volk, M. (2010): MT-based sentence alignment for OCR-generated parallel texts. In Proceedings of AMTA. Denver.

Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., Ruef, B. (2010a): Challenges in building a multilingual alpine heritage corpus. In Proceedings of the Seventh international conference on Language Resources and Evaluation (LREC).

Volk, M., Marek, T., Sennrich, R. (2010b): Reducing OCR errors by combining two OCR systems. In Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010).